

RD-A128 561

THE USE OF INFLUENCE FUNCTIONS FOR OUTLIER DETECTION
AND DATA EDITING. (U) AEROSPACE CORP EL SEGUNDO CA
GUIDANCE AND CONTROL DIV M R CHERNICK ET AL. 15 SEP 82
TR-0002(9990)-4 SD-TR-82-65

1/1

UNCLASSIFIED

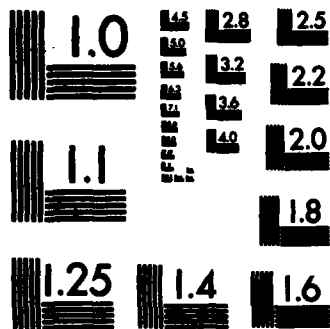
F/G 12/1

NL

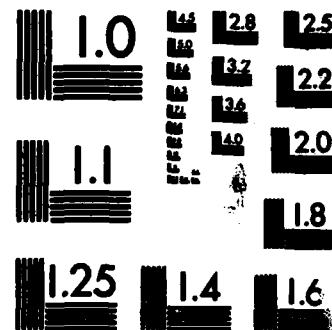
END

FILED

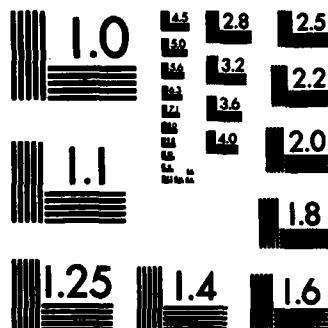
DTN



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



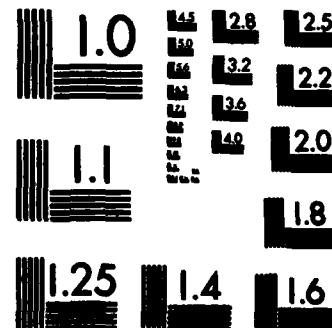
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A120561

DTIC F

Prepared for
SPACE DIVISION
AIR FORCE SYSTEMS COMMAND
Los Angeles Air Force Station
P.O. Box 92960, Worldway Postal Center
Los Angeles, Calif. 90009

DTIC
SELECTED
OCT 21 1982

A

82 10 21 002

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SD-TR-82-65	2. GOVT ACCESSION NO. AD-A120561	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE USE OF INFLUENCE FUNCTIONS FOR OUTLIER DETECTION AND DATA EDITING		5. TYPE OF REPORT & PERIOD COVERED
7. AUTHOR(s) M. R. Chernick and V. K. Murthy		6. PERFORMING ORG. REPORT NUMBER TR-0082(9990)-4
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Aerospace Corporation El Segundo, CA 90245		8. CONTRACT OR GRANT NUMBER(s) F04701-81-C-0082
11. CONTROLLING OFFICE NAME AND ADDRESS Space Division Air Force Systems Command Los Angeles, Calif. 90009		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE 15 Sep 82
		13. NUMBER OF PAGES 16
		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Influence functions Outliers Data editing Least squares estimation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Hampel's influence function (J. Amer. Statist. Assoc. 69, pp. 383-393) has been used in recent years for evaluating robust estimators, detecting outliers, computing asymptotic variances for estimators, and for hypothesis testing. In this report, the use of influence functions for various parameters is proposed, not only as a tool for outlier detection, but also as a method for replacing outliers and/or missing observations. This approach is illustrated on the estimation for the variance of a distribution and it is pointed out how the method can be applied in simple linear regression problems. This approach		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

19. KEY WORDS (Continued)

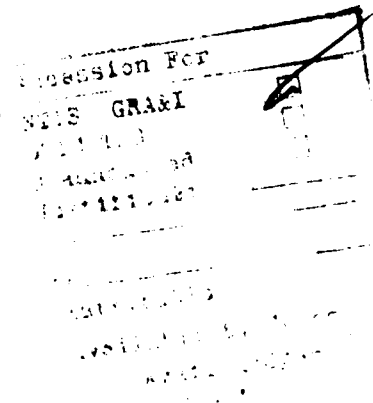
20. ABSTRACT (Continued)

can be extended to more complex regression type applications such as the estimation of the orbit parameters of a satellite. 

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

CONTENTS

1.	INTRODUCTION.....	3
2.	EXAMPLES.....	7
2.1	The Mean of a Distribution.....	7
2.2	The Variance of a Distribution.....	7
2.3	The Bivariate Coefficient of Correlation for a Bivariate Distribution Function.....	8
2.4	Slope and Intercept Parameters in a Simple Linear Regression.....	9
3.	OUTLIER DETECTION AND REPLACEMENT OF MISSING OBSERVATIONS.....	11
4.	AN EXAMPLE OF POWER PLANT DATA.....	17
	REFERENCES	19



A

1. INTRODUCTION

Hampel (Ref. 1) introduced the influence function as a tool for assessing robust estimators. In a multivariate population, an influence function can usually be defined for estimators of parameters. This influence function can be used to determine where in the n -dimensional space of observations the observed vector would have a large effect on the value of the estimator of the parameter.

For many parameters, the analytic form of the influence function can be derived (e.g., the mean, the variance, the bivariate correlation coefficient, and the multiple correlation coefficient). In other cases, it is difficult to obtain a closed-form expression for the influence function. In such cases, an empiric estimate may be useful.

Many agencies of the Federal government maintain large data bases and publish reports containing statistical information (e.g., the Bureau of the Census, the former Department of Energy, and the National Bureau of Standards). These agencies use outlier detection and data editing methods as quality control measures for their data bases. Also, the Department of Energy has had an extensive program for reviewing several of its data bases. In this data validation program, some new approaches for detecting outliers were introduced including the use of influence functions. (Chernick [Ref. 2])

At some of these agencies (particularly the Bureau of the Census) much research has been conducted on the replacement of observations (commonly called imputation). These techniques often rely on other related data to obtain a reasonable estimate as a replacement for the "bad" value. Also, in some cases, the method is designed so as to avoid inducing a large bias on the estimate of a particular parameter. Unfortunately, many of these data bases serve several purposes and a favorable procedure for one estimate may adversely affect estimates of other parameters which are important to different users of the data base. Consequently, influence functions can play an important role in the maintenance and validation of large data bases.

To give a formal definition, we point out that the influence function depends on the distribution function F of a random observation vector, the parameter of interest, which is commonly written $T(F)$, and the observation vector. The parameter is considered as a functional $T(F)$ of the distribution function F . The influence function is defined by the following equation whenever the limit on the right hand side exists:

$$I(F, T(F), \tilde{x}) = \lim_{\epsilon \rightarrow 0} \frac{T((1-\epsilon) F + \epsilon \delta_{\tilde{x}}) - T(F)}{\epsilon}$$

where ϵ is a positive real number, \tilde{x} is a point of interest in the observation space, and $\delta_{\tilde{x}}$ is the distribution function with all its probability mass concentrated at \tilde{x} .

The influence function is approximately equal in large samples to n times the difference between the estimator with an observation at \tilde{x} included and the estimator with the observation at \tilde{x} excluded where n is the sample size. This can be seen by replacing F with F_{n-1} (the empiric distribution function for a sample of size $n-1$) and approximating the limit as ϵ tends to 0 by replacing ϵ with $\frac{1}{n}$,

because $(\frac{n-1}{n})F_{n-1} + \frac{1}{n}\delta_{\tilde{x}} = F_n$, we get that the influence function is

approximately $n(T(F_n) - T(F_{n-1}))$ as claimed.

Given a sample of observations, the influence function for a parameter of interest may be estimated for each observation. An observation which has a very large estimated influence* will deserve particular attention, and we may be better off to discount it in our estimation procedure. If it is necessary for such an observation to be replaced, we may choose a value with small or zero influence on our estimator. The next section contains some common examples of this concept.

*What constitutes a large estimated influence depends on what assumptions are made about the underlying distribution.

In orbit determination problems, least-squares fitting methods are used to estimate orbit parameters based on data such as pseudo-range, delta range, and azimuth and elevation angle measurements. It is well known that the least-squares procedure leads to parameter estimates that can be very sensitive to outlying observations. Consequently, robust filtering or outlier rejection techniques sometimes need to be used to obtain good estimates of orbital elements on the basis of such data. The methodology proposed in this report can be used to arrive at more sophisticated outlier rejection and replacement techniques for the processing of these data.

The literature on influence functions is growing rapidly with new applications to estimation, hypothesis testing, and outlier detection appearing regularly. Chernick (Ref. 2) mentions an application to the validation of energy data and computes an influence function for multiple correlation in a special case. Chernick, Downing and Pike (Ref. 3) introduce an influence function matrix for the autocorrelation function which can be applied to detect outliers in time series. Reid (Ref. 4) determines an influence function for the Kaplan-Meier estimator of a survival curve and uses it to obtain the asymptotic variance of that estimator. Their use for hypothesis testing is proposed by Lambert (Ref. 5). Devlin, Gnanadesikan, and Kettenring (Ref. 6) illustrate the potential use of the influence function for detecting outlying multivariate observations.

2. EXAMPLES

2.1 THE MEAN OF A DISTRIBUTION

For a univariate distribution function F , the mean can be written as

$$T(F) = \int_{-\infty}^{\infty} y dF \text{ and the sample mean as } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = T(F_n),$$

where we assume $|T(F)| < \infty$ and $\{X_i\}_{i=1}^n$ are the n independent observations

and F_n is the empiric distribution function on the basis of these observations. In this case

$$\begin{aligned} I(F, T(F), x) &= \lim_{\epsilon \rightarrow 0} \frac{(1-\epsilon) \int_{-\infty}^{\infty} y dF + \epsilon x - \int_{-\infty}^{\infty} y dF}{\epsilon} \\ &= x - \int_{-\infty}^{\infty} y dF = T(F) = x - \mu \end{aligned} \quad (1)$$

Replacing μ by \bar{X} , we obtain a sample estimate for I

$$\hat{I} = x - \bar{X}.$$

This estimate of I is unbiased and consistent. Large values of \hat{I} correspond to observations which are say 2 or 3 standard deviations away from the mean. So the influence function for the mean is equivalent to a 3 sigma outlier rejection rule. If the observations have a normal distribution, the probability of the influence function estimate exceeding 3 standard deviations is less than 0.01.

2.2 THE VARIANCE OF DISTRIBUTION

The variance, σ^2 , of a univariate distribution function F can be written

$$T(F) = \int_{-\infty}^{\infty} (y-\mu)^2 dF \text{ and the sample variance } s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = T(F_n).$$

Again, this is meaningful only if we assume $T(F) < \infty$. The influence function is given by $I(F, T(F), x) = (x - \mu)^2 - T(F) = (x - \mu)^2 - \sigma^2$. (2)

A sample estimate could be

$$\hat{I} = (x - \bar{X})^2 - s^2.$$

This estimate is consistent but has bias σ^2/n . Note that large positive values of I again correspond to X s which are 3 or more standard deviations away from the mean. However, interestingly, the largest negative influence occurs at $x = \bar{X}$.

2.3 THE BIVARIATE COEFFICIENT OF CORRELATION FOR A BIVARIATE DISTRIBUTION FUNCTION

Here $\underline{x} = (x_1, x_2)$ is a two-dimensional vector,

$$T(F) = \frac{-\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 dF - \int_{-\infty}^{\infty} x_1 dF_1 \int_{-\infty}^{\infty} x_2 dF_2}{\left(\int_{-\infty}^{\infty} x_1^2 dF_1 - \left(\int_{-\infty}^{\infty} x_1 dF_1 \right)^2 \right) \left(\int_{-\infty}^{\infty} x_2^2 dF_2 - \left(\int_{-\infty}^{\infty} x_2 dF_2 \right)^2 \right)}$$

where F is the bivariate distribution defined by

$$F(x_1, x_2) = P[X_1 \leq x_1, X_2 \leq x_2]$$

and $F_1(x_1) = F(x_1, \infty)$, $F_2(x_2) = F(\infty, x_2)$.

In this case, the influence function is

$$I(F, T(F), \underline{x}) = y_1 y_2 - \rho \left(\frac{y_1^2}{2} + \frac{y_2^2}{2} \right) \quad (3)$$

where $y_1 = \frac{x_1 - \mu_1}{\sigma_1}$ and $y_2 = \frac{x_2 - \mu_2}{\sigma_2}$, μ_1 is the mean of F_1 ,

σ_1^2 is the variance of F_1 , μ_2 is the mean of F_2 , σ_2^2 is the variance

of F_2 , and $\rho = T(F)$. An estimate \hat{I} can be obtained by replacing ρ , μ_1 , μ_2 , σ_1 , and σ_2 with some estimates in Eq. (3). For a derivation of Eq. (3) see Chernick (Ref. 2). Gnanadesikan (Ref. 7) points out that for bivariate normal data, the estimated influence function for the Z transformation of the correlation coefficient has approximately a product standard normal distribution. This distribution can be used to determine significantly large values for I .

2.4 SLOPE AND INTERCEPT PARAMETERS IN A SIMPLE LINEAR REGRESSION

Here we assume

$$E(Y|X = x) = \alpha + \beta x \quad (4)$$

and let $\mu_y = E(y)$, $\mu_x = E(x)$, $\sigma_y^2 = \text{Var } Y$ and $\sigma_x^2 = \text{Var } X$. X and Y have a bivariate distribution function with finite second moments

Let $\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$. We have from Eq. (4)

$$\beta = \rho \frac{\sigma_y}{\sigma_x} \quad (5)$$

and

$$\alpha = \mu_y - \beta \mu_x. \quad (6)$$

Therefore

$$\begin{aligned} I(F, \beta, (x, y)) &= \frac{\sigma_y}{\sigma_x} I(F, \rho, (x, y)) + \rho \left(\frac{I(F, \sigma_y, y)}{\sigma_x} - \sigma_y \frac{I(F, \sigma_x^2, x)}{\sigma_x^2} \right) \\ &= \frac{\sigma_y}{\sigma_x} I(F, \rho, (x, y)) + \rho \left(\frac{I(F, \sigma_y^2, y)}{2\sigma_y \sigma_x} - \sigma_y \frac{I(F, \sigma_x^2, x)}{2\sigma_x^3} \right) \end{aligned} \quad (7)$$

and from Eqs. (2) and (3)

$$I(F, \beta, [x, y]) = \frac{\sigma_y}{\sigma_x} \left(\left(\frac{x - \mu_x}{\sigma_x} \right) \left(\frac{y - \mu_y}{\sigma_y} \right) \frac{-\rho}{2} \left\{ \left(\frac{x - \mu_x}{\sigma_x} \right)^2 + \left(\frac{y - \mu_y}{\sigma_y} \right)^2 \right\} \right) + \rho \left[\frac{[(y - \mu_y)^2 - \sigma_y^2]}{2 \sigma_y \sigma_x} - \frac{[(x - \mu_x)^2 - \sigma_x^2] \sigma_y}{2 \sigma_x^3} \right] \quad (8)$$

From Eq. (6) we get

$$I(F, \alpha, (x, y)) = (y - \mu_y) - \beta(x - \mu_x) - \mu_x I(F, \beta, (x, y)) \quad (9)$$

From Eqs. (8) and (9) we see that these influence functions can be estimated by obtaining sample estimates of $\mu_x, \mu_y, \sigma_x, \sigma_y$, and ρ . For the regression parameters we see that the influence function depends on the same parameters as for the correlation coefficient.

Techniques for determining influential observations and leverage points in regression problems are given in Belsley, Kuh, and Welsch (Ref. 8). Because the classical approach to orbit determination involves a linearization which leads to solving a large regression problem (i.e., estimating six or more parameters) the techniques given by Belsley, Kuh, and Welsch could be useful. Also, an influence function for the regression parameters could be calculated similarly to the calculation illustrated here for the simple linear regression.

In each of the four examples given in this section, the estimator \hat{I} will be consistent for I if consistent estimates of the unknown parameters are used. Also, if the maximum likelihood estimates are used for the unknown parameters the estimator \hat{I} will be a maximum likelihood estimator of I .

3. OUTLIER DETECTION AND REPLACEMENT OF MISSING OBSERVATIONS

In Gnanadesikan (Ref. 7), it is shown how contours of constant influence based on Eq. (3) can be used as a graphical tool for detecting outliers with respect to bivariate correlation. Chernick, Downing, and Pike (Ref. 3) use influence function estimates for the autocorrelation function of a time series to determine outliers.

Here we propose the use of the influence function to determine observation values for replacing outliers or for replacing missing observations. Observations with unduly high influence should be replaced by values which have little or no influence on the estimated parameter or parameters. The philosophy is that if an observation needs to be replaced and no additional information is available about what the correct value should be, then one should choose a value that does not influence estimates of importance to the users of the data. All this assumes that there is a need to replace the outlier or to fill in a missing observation.

We shall now illustrate this approach for the case when the estimate of interest is the population variance (i.e., example Eq. (2)).

Here we assume that we have a sample of size n and we are concerned that we might have one or two outliers. By rearranging the numbering of the observations, if one outlier is detected, we assume for notational convenience that it is the n th observation.

Suppose in the case of one outlier that we observed a value

$X_n = \bar{X}_{n-1} + 3 S_{n-1}^2$. The influence function estimate for both the mean and

the variance at X_n will be large indicating that the observation is an outlier, we shall replace X_n with \hat{X}_n .

Let

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n-1} X_i + \frac{1}{n} \hat{X}_n \quad (10)$$

$$\bar{x}_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i,$$

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x}_{n-1})^2$$

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n-1} (x_i - \bar{x}_n)^2 + \frac{1}{n} (\hat{x}_n - \bar{x}_n)^2, \quad (11)$$

where \hat{x}_n is our choice for a replacement to x_n . From Eqs. (10) and (11) we get

$$\bar{x}_n - \bar{x}_{n-1} = \frac{\hat{x}_n - \bar{x}_{n-1}}{n} \quad (12)$$

and

$$s_n^2 - s_{n-1}^2 = \frac{-s_{n-1}^2}{n} + \frac{n-1}{n^2} (\hat{x}_n - \bar{x}_{n-1})^2. \quad (13)$$

Ideally, we would like to choose \hat{x}_n so that $\bar{x}_n = \bar{x}_{n-1}$ and $s_n^2 = s_{n-1}^2$, however,

we cannot quite do this. If we choose $\hat{x}_n = \bar{x}_{n-1}$, then $\bar{x}_n = \bar{x}_{n-1}$. However,

choosing $\hat{x}_n = \bar{x}_{n-1}$ tends to make $s_n^2 < s_{n-1}^2$. In fact, from Eq. (13) we see

that $s_n^2 - s_{n-1}^2 = \frac{-s_{n-1}^2}{n}$. On the other hand, Eq. (2) tells us that

$I(F, T(F), x) = 0$ if $(x - \mu)^2 = \sigma^2$ or $x = \mu \pm \sigma$. Consequently, one would

suspect that the choice $\hat{x}_n = \bar{x}_{n-1} + s_{n-1}$ or $\hat{x}_n = \bar{x}_{n-1} - s_{n-1}$ would have a

smaller influence on the estimate of σ^2 . This suspicion is borne out by the

fact that for either choice $S_n^2 - S_{n-1}^2 = \frac{-S_{n-1}^2}{n}$. The price paid for reducing

the influence on the variance is an influence on the mean. When

$$\hat{X}_n = \bar{X}_{n-1} + S_{n-1}, \quad \bar{X}_n - \bar{X}_{n-1} = \frac{S_{n-1}}{n} \text{ and when}$$

$$\hat{X}_n = \bar{X}_{n-1} - S_{n-1}, \quad \bar{X}_n - \bar{X}_{n-1} = \frac{-S_{n-1}}{n}. \text{ A slight modification of this}$$

replacement for X_n leads to a zero change in the estimate of σ^2 . If we solve

the equation $S_n^2 = S_{n-1}^2$ for \hat{X}_n , we find $\hat{X}_n = \bar{X}_{n-1} \pm \sqrt{\frac{n}{n-1}} S_{n-1}$. In the

case $\hat{X}_n = \bar{X}_{n-1} + \sqrt{\frac{n}{n-1}} S_{n-1}$ we have $\bar{X}_n - \bar{X}_{n-1} = \frac{S_{n-1}}{\sqrt{n(n-1)}}$ and when

$$\hat{X}_n = \bar{X}_{n-1} - \sqrt{\frac{n}{n-1}} S_{n-1} \text{ we have } \bar{X}_n - \bar{X}_{n-1} = \frac{-S_{n-1}}{\sqrt{n(n-1)}}. \text{ We shall now consider}$$

the case of two outliers.

Theorem 1

In the case of two outliers, say X_{n-1} and X_n , we can choose

$$\hat{X}_{n-1} = \bar{X}_{n-2} + S_{n-2}$$

and

$$\hat{X}_n = \bar{X}_{n-2} - S_{n-2}$$

where

$$\bar{X}_{n-2} = \frac{1}{n-2} \sum_{i=1}^{n-2} X_i$$

and

$$S_{n-2}^2 = \frac{1}{n-2} \sum_{i=1}^{n-2} (X_i - \bar{X}_{n-2})^2.$$

In this case we have $\bar{X}_n - \bar{X}_{n-2} = 0$ and $S_n^2 - S_{n-2}^2 = 0$.

proof:
$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n-2} X_i + (\hat{X}_n + \hat{X}_{n-1})/n$$

$$= \frac{(n-2)}{n} \bar{X}_{n-2} + (\hat{X}_n + \hat{X}_{n-1})/n.$$

Our choice of $\hat{X}_n = \hat{X}_{n-1}$ makes $(\hat{X}_n + \hat{X}_{n-1})/n = 2\bar{X}_{n-2}/n$. Consequently,

$$\bar{X}_n = \bar{X}_{n-2}. \quad (14)$$

From Eq. (14), we see that

$$\begin{aligned} S_n^2 &= \frac{1}{n} \sum_{i=1}^{n-2} (X_i - \bar{X}_{n-2})^2 + (\hat{X}_{n-1} - \bar{X}_{n-2})^2/n \\ &+ (\hat{X}_n - \bar{X}_{n-2})^2/n = \frac{n-2}{n} S_{n-2}^2 + \frac{2 S_{n-2}^2}{n} \\ &= S_{n-2}^2. \end{aligned} \quad (15)$$

We note that as long as we choose $\hat{X}_n = 2\bar{X}_{n-2} - \hat{X}_{n-1}$ Eq. (14) will be satisfied. But in order to also obtain Eq. (15), we must have

$\hat{X}_{n-1} = \bar{X}_{n-2} \pm S_{n-2}$. We are fortunate that in this situation, because of symmetry, we can find replacement values which leave both the mean and variance estimates unchanged.

In the case of one outlier, we cannot do this. If estimating the mean is important, and we do not care about the estimate of variance, one should choose $\hat{X}_n = \bar{X}_{n-1}$. On the other hand, if the estimate of variance is much more important, one should use $\hat{X}_n = \bar{X}_{n-1} + S_{n-1}$ or $\hat{X}_n = \bar{X}_{n-1} - S_{n-1}$, with the choice among these two estimates dictated by whether X_n was larger than \bar{X}_{n-1} or not (in the case when the outlying observation X_n is known). If both parameters are important in the estimation, a compromise choice for \hat{X}_n should be chosen perhaps by minimizing a weighted average of the absolute value (or square) of the estimated influence functions.

The approach of minimizing a weighted average of the absolute value (or square) of the influence function can be generalized to the case of several parameters. Observations will be declared outliers if they have an unduly large influence on any of the important parameters and they will be replaced by values that minimize their average absolute or average square influence.

The methodology described here for the variance could also be applied in the case of the correlation coefficient or the regression parameters or for combinations of these parameters.

4. AN EXAMPLE OF POWER PLANT DATA

The Department of Energy collects monthly data on fuel consumption and electricity generation for all utilities and some industrial plants. For two particular plants, three years of monthly data were analyzed and outliers were found using the influence function for bivariate correlation (Chernick [Ref. 2]). Subsequently, this data was used to illustrate a new time series technique for detecting outliers based on influence functions (Chernick, et al. [Ref. 3]). Table 1 presents the 36 values of the consumption data for one of the plants. This table also includes values for the influence function estimates for the mean and variance of the sample at each observation point. Notice that observation no. 23 has the largest influence on both the mean and variance. In computing the influence function for the correlation between consumption and generation, observation no. 23 also stood out. Note that for the variance, the observations closest to the sample mean have the largest negative influence although these observations have the smallest influence on the mean.

Assuming that one wanted to impute a value to observation no. 23, we would choose the value 9.3 if we wanted to leave the mean unchanged. If we wanted to leave the variance unchanged, we use either $9.3 + 0.38 = 9.68$ or $9.3 - 0.38 = 8.92$. Because the consumption and generation data are highly correlated and the generation data was not suspect in this case, an estimate for the consumption data on the basis of regression or on the influence function for bivariate correlation would be more appropriate in this application.

Table 1. Influence Function Estimates for Power Plant Data

<u>Observation Number</u>	<u>Consumption</u>	<u>Influence Functions</u>	
		<u>Mean</u>	<u>Variance</u>
1	71	59.4	3167.3
2	4	-7.6	-303.3
3	3	-8.6	-287.1
4	5	-6.6	-317.5
5	4	-7.6	-303.3
6	3	-8.6	-287.1
7	48	36.4	963.9
8	11	-0.6	-360.7
9	11	-0.6	-360.7
10	5	-6.6	-317.5
11	4	-7.6	-303.3
12	16	4.4	-341.7
13	3	-8.6	-287.1
14	4	-7.6	-303.3
15	6	-5.6	-329.7
16	15	3.4	-349.5
17	5	-6.6	-317.5
18	3	-8.6	-287.1
19	4	-7.6	-303.3
20	13	1.4	-359.1
21	6	-5.6	-329.7
22	5	-6.6	-317.5
23	93	81.4	6264.9
24	4	-7.6	-303.3
25	8	-3.6	-348.1
26	4	-7.6	-303.3
27	5	-6.6	-317.5
28	5	-6.6	-317.5
29	4	-7.6	-303.3
30	6	-5.6	-329.7
31	19	7.5	-306.3
32	4	-7.6	-303.3
33	3	-8.6	-287.1
34	3	-8.6	-287.1
35	5	-6.6	-317.5
36	4	-7.6	-303.3

REFERENCES

1. F. R. Hampel, "The Influence Curve and Its Role in Robust Estimation," J. Amer. Stat. Assoc. 69, pp. 383-393 (1974).
2. M. R. Chernick, "The Influence Function and its Application to Data Validation," ORNL/TM 6871, Oak Ridge National Laboratory, Oak Ridge, Tennessee (1979).
3. M. R. Chernick, D. J. Downing, and D. H. Pike, "Detecting Outliers in Energy Time Series Data," Proceedings of the Business and Economic Statistics Section, American Statistical Association, pp. 101-106 (1980).
4. N. Reid, "Influence Functions for Censored Data," The Annals of Statistics 9, pp. 78-92 (1981).
5. D. Lambert, "Influence Functions for Testing," J. Amer. Stat. Assoc. 76, pp. 649-657 (1981).
6. S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring, "Robust Estimation and Outlier Detection with Correlation Coefficients," Biometrika 62, pp. 531-545 (1975).
7. R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, Wiley, New York (1977).
8. D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics, Wiley, New York (1980).
9. C. L. Mallows, "On Some Topics in Robustness," unpublished paper (1974).

END

FILMED